PSI: A Benchmark for Human Interpretation and Response in Traffic Interactions

Taotao Jing[1] **Tina Chen**[2] **Renran Tian**[3]* **Yaobin Chen**[2] **Joshua Domeyer**[4]
Heishiro Toyoda[4] **Rini Sherony**[4] **Zhengming Ding**[1]
[1]Tulane University [2]Purdue University
[3]North Carolina State University [4]Toyota Motor North America
{tjing, zding1}@tulane.edu cchen.tina@gmail.com rtian2@ncsu.edu chen62@purdue.edu
{joshua.domeyer, rini.sherony}@toyota.com heishiro.toyoda.tmc@tri.global

Abstract

Accurately modeling pedestrian intention and understanding driver decision-making processes are critical for the development of safe and socially aware autonomous driving systems. However, existing datasets primarily emphasize observable behavior, offering limited insight into the underlying causal reasoning that informs human interpretation and response during traffic interactions. To address this gap, we introduce PSI, a benchmark dataset that captures the dynamic evolution of pedestrian crossing intentions from the driver's perspective, enriched with human-annotated textual explanations that reflect the reasoning behind intention estimation and driving decision making. These annotations offer a unique foundation for developing and benchmarking models that combine predictive performance with interpretable and human-aligned reasoning. PSI supports standardized tasks and evaluation protocols across multiple dimensions, including pedestrian intention prediction, driver decision modeling, reasoning generation, and trajectory forecasting and more. By enabling causal and interpretable evaluation, PSI advances research toward autonomous systems that can reason, act, and explain in alignment with human cognitive processes. The dataset is publicly available at `http://situated-intent.net/pedestrian_dataset/`.

1 Introduction

As automation capabilities increase, autonomous vehicles (AVs) can perform their driving roles more easily in highway and freeway contexts. However, the complexity of mixed urban traffic, especially with pedestrians, remains challenging [14, 6]. Interactions with pedestrians are safety-critical and affect traffic efficiency and user attitudes [16]. To operate around pedestrians, most AVs make decisions in responding to their trajectories and optimize the ego motions to avoid crashes [34, 10]. Vehicle dynamics are often calculated based on the sensing inputs of motion parameters from the surrounding objects [54]. To support vehicle decision-making during pedestrian encounters, learning-based approaches have been applied to predict pedestrian trajectories and actions [4, 34, 25], with an increasing number of algorithms and benchmark datasets published recently [11, 58, 55, 24, 21]. However, pedestrian behavior prediction models have reduced performance for long duration predictions and still struggle to handle sudden action changes [7, 28, 42].

Driving is not only a technical skill but also a social skill [50, 41]. This perspective emphasizes the importance of social factors like the goals and intentions of different road users and the role of personal experiences in understanding and planning behaviors during interactions. Similar to human

*Corresponding Author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

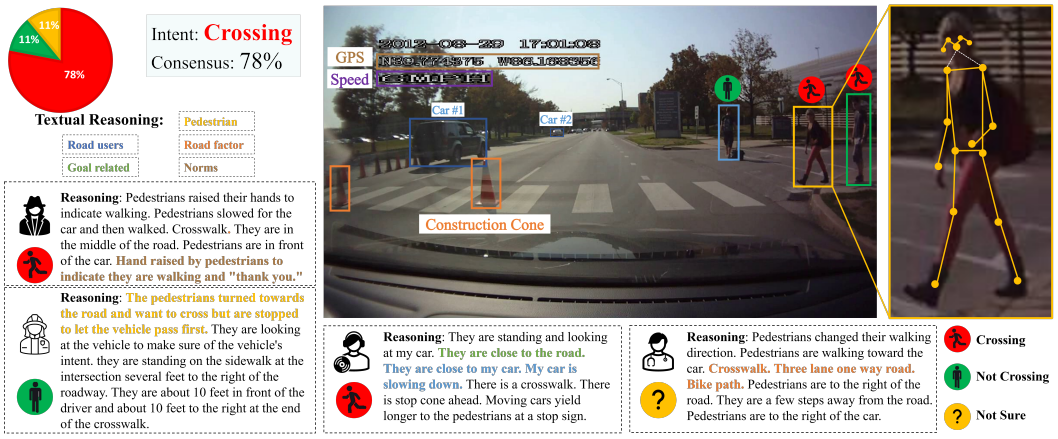


Figure 1: Annotation examples: (1) bounding boxes and postures of pedestrians, crossing intent and textual reasoning from multiple annotators, road users, street structure (traffic lights, stop signs, construction cones, etc.), GPS, and speed; (2) textual reasoning of the intent estimation with different topics (pedestrians, road users, road factor, goal-related, social norms).

drivers, AVs need to continuously coordinate with other human road users by exchanging intentions through communication [41]. Thus, pedestrian intention prediction has been widely discussed in the literature, with many earlier studies using trajectories and actions (e.g., walking) as surrogates of actual pedestrian intention [21, 62]. Considering that human intention reflects the degree of motivations to influence or act certain behaviors [1] and there is an intention-behavior gap [45], i.e., people do not or cannot always do the things they intend to do, recent research recognizes the importance of directly studying pedestrian intentions in addition to the observable behaviors. It is expected that the prediction of intentions can advance the timing of accurately predicting actions, while also improving the accuracy of predicting long-term behavior and behavioral changing points.

One pioneering study [43] applied a crowd-sourcing method to label pedestrian crossing intentions. The authors asked data annotators to watch pre-recorded pedestrian videos and estimate the pedestrians' intentions to cross the street at one pre-determined critical frame for each video. Although the datasets support many models [52, 59], there are several limitations that need to be addressed, including unclear intention definition, missing intention changes, label subjectivity, and lack of reasoning and background knowledge, which will be discussed in details in section 1.1.

1.1 Related Datasets

In the last few years, there has been a sharp increase in the number of datasets released for pedestrian behaviors and automated vehicles research. These datasets are typically collected in major urban environments with pedestrian-centric annotations. Several datasets focus on or have labels related to pedestrian behaviors. The earlier data set, JAAD [33], annotated pedestrian actions to study their crossing behavior with scene attributes also heavily annotated. STIP [36] contains pedestrian crossing action annotations, but lack additional attributes other than tracked bounding boxes. TITAN [48] classifies pedestrian actions into 43 sub-categories to try to explain their behavior. PedX [32] focuses on 3D human pose annotations to explain pedestrian behavior. Differently, Euro-PVI [5] is a dataset of pedestrian and bicyclist trajectories that encompasses diverse vehicle-pedestrian (and bicyclist) interactions. VIENA2 [2] gathers a large-scale dataset covering various driving conditions and environments obtained using the GTA-V video game, and the data are annotated with a total of 25 distinct action classes for anticipatory driving action.

For explainable AV datasets, normally there are both visual inputs and text annotations. Berkeley Deep Drive eXplanation (BDD-X) dataset [31] is a subset of the BDD dataset [56] with action description and justification for all the driving events. BDD Object Induced Actions (BDD-OIA) dataset [57] is a subset of BDD100K [60]. To increase scene diversity, these videos were selected under various weather conditions and times of the day. This resulted in 22,924 five-second (5-s) video clips, which were annotated on MTurk for the 4 actions and 21 explanations. However, none of the existing datasets explains pedestrian crossing intents in naturalistic scenes.

Few earlier datasets address pedestrian intentions. The PIE [43] dataset directly labels pedestrian crossing intentions using crowd-sourcing workers. Several other studies, like the LOKI [21] dataset, use future pedestrian actions to surrogate current intentions. However, these datasets have some key limitations. First, current annotations fail to distinguish between lower-level and higher-level intentions. Psychological literature defines two types of human intentions: future-directed intentions, which support planning and the formation of sub-intentions, and present-directed intentions, which directly lead to actions [8, 13]. In pedestrian behavior, future-directed intentions refer to the general plan to cross a road—useful for route planning but not necessarily indicative of immediate behavior—while present-directed intentions influence real-time actions such as stopping or crossing. Most existing datasets conflate these two, resulting in mixed or conflicting intention labels and predictions. Second, while future-directed intentions typically remain stable during an encounter, present-directed intentions are highly dynamic and influenced by moment-to-moment contextual changes. Ethnographic studies have documented behavioral shifts during vehicle-pedestrian negotiations [15, 46], but prior datasets only provide annotations at isolated critical frames, neglecting the evolving nature of real-time decision-making. This temporal gap introduces uncertainty in model training and evaluation when intention labels are assumed to hold across entire encounters.

More importantly, crowd-sourced intention annotations often suffer from subjectivity. Disagreements among annotators are common in subjective labeling tasks [29], and divergent estimations of pedestrian intentions have been noted in published datasets [43]. Although soft labels [37, 53] and annotation aggregation methods [47, 51] can mitigate this issue, robust modeling of annotation uncertainty requires datasets with a large, demographically balanced pool of annotators—an element lacking in current datasets. Finally, existing datasets lack insight into the reasoning behind intention estimation. Pedestrian behavior is influenced by a rich array of contextual and behavioral cues [44, 20], yet most datasets exclude verbal reasoning. As foundation models and vision-language models (VLMs) have advanced capabilities in generalization, few-shot learning [9], and in-context learning [17], incorporating human reasoning is critical to improve AI understanding and explainability [27, 30]. Verbal rationales from human drivers not only offer background knowledge but also serve as valuable resources for improving interpretability in autonomous systems [31, 61].

1.2 Objectives and Contributions

This study introduces a new dataset to measure **Pedestrian Situated Intent (PSI)**, as illustrated in Figure 1, defined as a pedestrian's intention to cross in front of the ego vehicle within a specific contextual interaction. PSI refers to the pedestrian's intention to cross the ego vehicle's path at a particular location before the vehicle arrives at that point. This definition focuses on immediate, present-directed crossing actions from the AV's perspective, highlighting the associated risks and urgency. Unlike broader definitions of pedestrian intent as a general desire to cross the road, PSI provides more temporal and spatial boundaries, making it especially relevant for AV decision-making.

Table 1: Comparison of popular pedestrian behavior datasets in the AV research area. *Number of annotated frames only include 2D annotations. †Number of classes used for annotating objects in the dataset. ‡Number of frames manually annotated. **Dataset contains LiDAR but we do not consider it here. "Disagr." denotes the disagreement across multiple observers for the same case.

			Visual Annotations			Pedestrian Intent			Driving Decision		
Dataset	**Year**	**#Frames***	**B-Boxes**	**Pose**	**Class†**	**Intent**	**Disagr.**	**Reason**	**Drive**	**Disagr.**	**Reason**
Ours (PSI)	2022	79k	✔	✔	10	987k	✔	✔	✔	✔	✔
LOKI [21]	2021	8k	✔	✗	8	28k	✗	✗	✗	✗	✗
STIP [36]	2020	110k‡	✔	✗	1	✗	✗	✗	✗	✗	✗
TITAN [38]	2020	75k	✔	✗	3	✗	✗	✗	✗	✗	✗
PIE [43]	2019	293k	✔	✗	6	27k	✗	✗	✗	✗	✗
PedX** [32]	2019	5k	✗	✔	1	✗	✗	✗	✗	✗	✗
JAAD [33]	2016	82k	✔	✗	3	✗	✗	✗	✗	✗	✗
BDD-100K [60]	2018	100k	✔	✗	10	✗	✗	✗	✗	✗	✗

With this definition, PSI represents a temporally dynamic mental state that evolves with interaction context which is not captured in existing benchmark datasets. The new PSI dataset addresses key challenges in intention annotation, including subjectivity, human bias, reasoning background, and algorithm explainability, by increasing annotator diversity and collecting textual reasoning behind

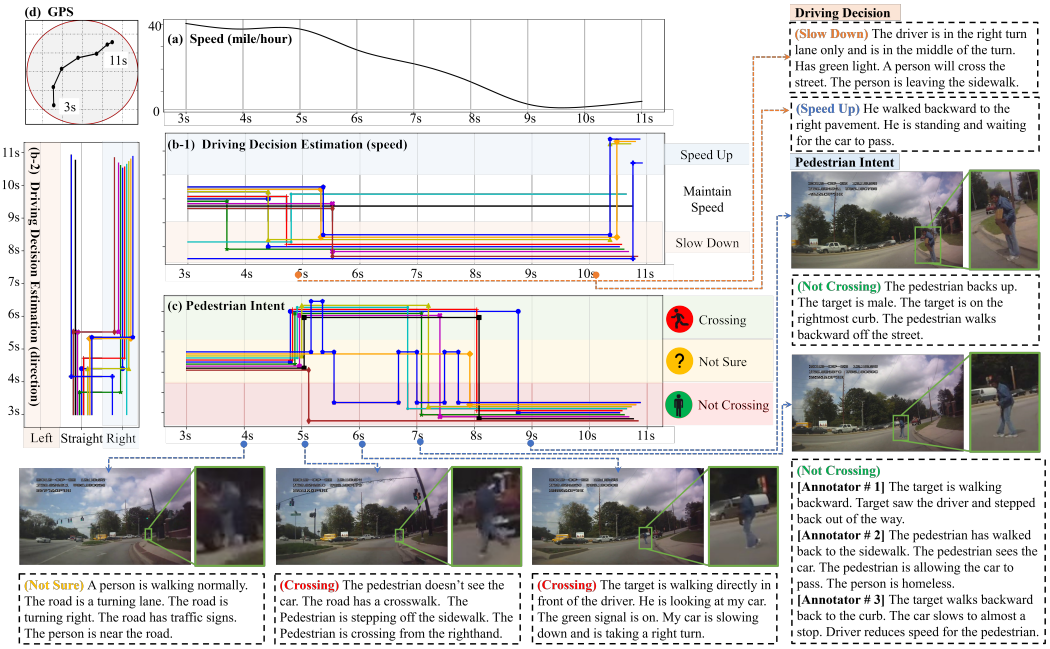


Figure 2: The example shows dynamic changes in estimated pedestrian situated intents and driving decisions (speed and direction), as well as reasoning descriptions from multiple annotators. Vehicle speed and GPS position are also included. The colored lines represent annotations from different annotators for the same case. Each annotator was asked to estimate the changes of the pedestrian's crossing intent and justify their reasoning, as well as their driving decisions (direction and speed). This dynamic estimation process reflects the interaction between pedestrians and drivers, leading to consensus or disagreement among annotators. At times, all annotators agree on the same estimation. Conversely, there may be two or three opinions when there is disagreement. This diversity in the PSI dataset enhances the robustness of the annotations and provides a more comprehensive understanding of the scene.

intention estimates. Intention annotations and reasoning were gathered from up to 24 annotators for each scene, who were drawn from varied age groups with balanced gender representation. The collected reasoning texts serve as a valuable knowledge base for AI decision-making. Additionally, the PSI dataset includes high-quality manual annotations of common visual features, including object detection and categorization, object tracking, and pedestrian postures. These rich annotations support a wide range of computer vision tasks using traditional learning techniques or most recent in-context learning, prompt engineering, and finetuning of foundation models.

We compare the new PSI dataset with several pedestrian-related dataset and show some key attributes as Table 1. From the comparison, we observe that PSI dataset is the first dataset annotating the pedestrian crossing intention and driving decision together *continuously* with reasoning descriptions and inter-driver disagreements (soft labels). It is noteworthy that the intention definition in the PSI dataset is present-directed which is different from all other crossing intention definitions [43, 33]. Because the PSI intention is annotated continuously across all the frames, the dataset has a much larger number of frames with intention annotations. These manually annotated frames provide more reliable intention labels for any sequences captured in the data, which are more accurate than the common practice of extending the intention annotation for one critical frame towards all the frames in the current pedestrian intention prediction research. To our best knowledge, disagreements among annotators and textual reasoning are explored for the first time in pedestrian scene understanding.

2 The PSI dataset

2.1 Data Preparation and Unique Features

The PSI dataset contains 196 vehicle-pedestrian encounters with potential conflicts. Different from most pedestrian behavior benchmark datasets whose videos are continuously recorded in urban and downtown areas, the PSI dataset was randomly sampled from over 70,000 pedestrian encounters

captured in the TASI 110-car naturalistic driving study [48]. In the driving study, 110 drivers were recruited to install a data collection system in each of their own cars. For one whole year, the positions of and the driving scenes in front of the subject's cars were continuously recorded. From 1.4 million miles of driving data, more than 70,000 pedestrian encounters were identified across all cars. Multiple data annotators manually checked all these pedestrian encounters to identify potential conflict cases. The potential conflict is defined as that at a particular timestamp during the encounter, crashes would happen if both the car and the pedestrian keep their instantaneous speed and directions [48]. A total of more than 3,000 potential conflict encounters were labeled in the whole data set, from which the 196 encounters were randomly selected for the PSI dataset. Each case is 15 seconds with scene videos at 30 fps, GPS coordinates at 1 fps, and vehicle speed at 1 fps.

Compared with other benchmark datasets, the PSI cases have the following unique features: **a**) The cases are more representative of the driving experiences of normal drivers in different road and environmental situations across the period of a whole year; **b**) Every case has potential conflicts from the human driver's perspective, indicating the existence of vehicle-pedestrian interactions and ensuring the focused pedestrians are relevant to the driving decision-making; **c**) All the cases contain longer and more complicated pedestrian behavior and action changes, making the prediction tasks more challenging and realistic; and **d**) The driving behaviors reflect more than 100 normal drivers to represent human driving decisions and corresponding pedestrian responses in the natural road environment. Because of all these features, the PSI dataset is a more challenging and realistic dataset to develop and test pedestrian intentions and behaviors to better support AVs.

2.2 Visual Annotations

Table 2: List of labels for visual annotations.

Bounding Box Labels	
Person	pedestrian, rider
Vehicle	car, bus, bicycle, semi-truck, motorcycle
Road	traffic sign, traffic light, construction cone

There are two types of annotations in PSI: visual and cognitive annotations. Visual annotations include bounding boxes and pedestrian pose estimation. Bounding boxes and poses are all tracking enabled, so each object and pose is tracked for the entirety of the video clip. We annotated 10 classes of traffic objects and agents for bounding boxes (pedestrian, rider, car, bus, bicycle, semi-truck, motorcycle, traffic sign, traffic light, construction cone). Table 2 lists the classes annotated with bounding boxes. For the 196 video clips, there are 79,837 frames annotated with lower-level visual annotations. In total, there are $682,378$ bounding box annotations for traffic objects and agents, along with $71,259$ unique pose annotations for 391 pedestrians, using the MS COCO format [35]. The pose annotations comprise 17 different keypoints (classes) annotated with three values (x, y, v), where x and y values denote the coordinates, and v indicates the visibility of the key point (visible / not visible). There are 373 pedestrians out of the 391 were identified as key pedestrians that also have cognitive annotations.

2.3 Cognitive Annotations

The cognitive annotations include pedestrian situated intention (PSI), human reasoning descriptions, and driver behavior during the interaction. (**a**) *Pedestrian Situated Intention* is segmented as polygonal chains (in Figure 2) representing the estimated pedestrian intentions to cross in front of the ego-vehicle. The intentions can switch from three states, namely "Cross", "Not Cross", and "Not Sure". Each polygonal curve is the time-based estimation from one human driver/annotator, classifying the pedestrian's intention into one of the three categories in every frame. The vertices form the segmentation boundaries when pedestrian intention changes. (**b**) *Driver Behavior* is annotated in response to the dynamics of the pedestrian intentions, including speed and direction changes along time. Similar to the PSI labels, frame-by-frame driver behavior labels from one annotator can be shown as one polygonal curve, representing the estimated driving speed changes among "Speed Up", "Maintain Speed", and "Slow Down" or direction changes among 'Left", "Straight", and "Right". (**c**) *Human Reasoning Descriptions* are collected as text explanations corresponding to all the vertices along all the pedestrian intention and driver behavior segmentation polygonal curves. These descriptions explain the most critical reasoning logic from human drivers/annotators during the preceding scene segments, and are collected via multiple prompts from five categories of pedestrian behavior influential factors [20] to stimulate in-deep thinking and explanations.

2.3.1 Annotator Statistics

74 data annotators were recruited in this study. Each annotator labels at least 45 encountering scene videos through a video experiment to obtain these labels. All the annotators (44 males and 30 females) have valid US driving licenses and are from 19 to 77 years old. The diverse backgrounds of annotators can ensure the representativeness of pedestrian intent estimation results and reasoning descriptions.

2.3.2 Intention Annotation Process

In order to obtain the cognitive annotations, a video experiment was conducted for 9 to 24 video annotators per video. To capture human cognition in estimating pedestrian intents and making driving decisions, we developed a browser-based data annotation platform that facilitated the annotation process. All 196 pedestrian encounter videos were presented to data annotators through the user interface. The videos would play and pause at the first frame when the pedestrian becomes visible, with the target pedestrian indicated by a red arrow. Annotators were tasked with estimating the pedestrian's intention to cross in front of the ego vehicle at that moment, based on preceding scenes, and providing a brief reasoning description to support their estimation. Several key functions include:

- A novel Point-and-Explain (PaE) UI that mimics how humans explain to one another. The interface allows annotators to refer to a visual objects and explain the related reasoning, and automatically links visual and verbal objects in the data.
- Multi-prompt reasoning prompts use open-ended and close-ended questions to stimulate the annotator to provide reasoning from different perspectives.

After the first frame, annotators need to estimate the pedestrian intent frame by frame until the end of the video. If the pedestrian's intention changed at a particular frame, a segmentation boundary was inserted, and a reasoning description was required to explain the situation. Each time a change in intention estimation was identified, reasoning was described in five aspects: pedestrian-related factors, other road users, goal-related factors, road factors, and social norms. To guide annotators and encourage more detailed reasoning descriptions, separate questions were provided as instructions.

Each video will be repeated twice to label pedestrian intents and driving decisions separately, following the same process. Details of the experiment platform and process can be found in [19].

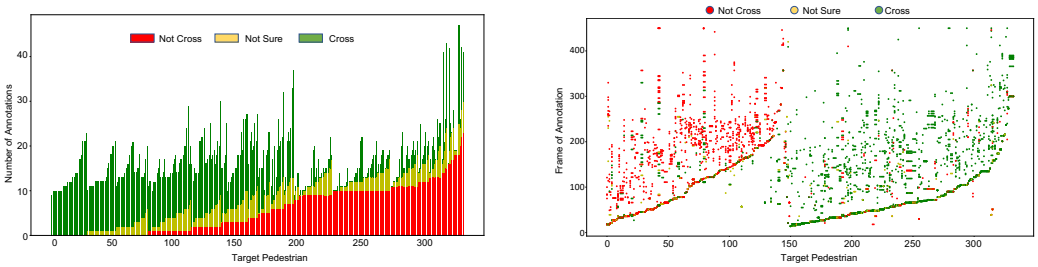
2.3.3 Cognitive Annotation Statistics

In the PSI dataset, 74 annotators have labeled the defined Pedestrian Situated Intents frame by frame to achieve a total of more than $987k$ intention estimations. There are $5,773$ Pedestrian Situated Intent segmentation boundaries in the entire data set across all the annotators and cases. On average, each case has 29.45 boundaries from all the annotators. Accordingly, there are $5,773$ segments in all the intention segmentation curves with the same number of reasoning descriptions. Each encounter has 9 to 24 polygonal segmentation curves from all the subjects. On average, each annotator has 1.2 segments for each encounter case in estimating the situated intent. The average length of the reasoning is 251.7 characters.

2.3.4 Annotation Discrepancies

Due to the involvement of multiple video annotators with diverse backgrounds, each encounter in the dataset exhibits both similarities and discrepancies in terms of intention estimation and reasoning descriptions. These similarities often reflect common sense and social norms, while the discrepancies primarily arise due to individual differences among the annotators. To quantify the level of agreement among the annotators or drivers in pedestrian intention estimation, we utilize the agreement score, which represents the highest percentage of annotators who agree on the same pedestrian intention at a particular timestamp. Conversely, the disagreement score is calculated as the complement of the agreement score (i.e., 1 minus the agreement score) for all frames and cases.

The analysis of the results reveals two significant findings. First, in over 58% of the encountered scenes, the average disagreement score throughout the entire duration remains below 20%. This indicates that in the majority of situations, most annotators are able to reach a consensus in estimating the defined PSI, reflecting a high degree of agreement. Second, more than 30% of the cases exhibit a maximum disagreement score exceeding 50%. This suggests that there are specific instances within many cases where more than half of the annotators fail to reach a consensus in their PSI estimations.



(a) Number of intention annotations obtained for each target pedestrian in the PSI dataset.

(b) Time distribution (frame) of pedestrians annotated.

Figure 3: Overview of intention annotation statistics in the PSI dataset.

These findings highlight the inherent complexity of real-world driving scenarios and emphasize the necessity of involving a large number of annotators to capture diverse perspectives. To support algorithm development and evaluation, the PSI dataset includes the original (soft) labels provided by all the different annotators. This facilitates further research and allows researchers to focus on scenes with higher or lower agreement scores, which can serve as indicators of task difficulties.

2.3.5 Intention Annotation Distribution

To explore the distribution of intention annotations in the PSI dataset, we present the number of intention annotations for each target pedestrian in Figure 3(a). It should be noted that certain videos in the dataset contain multiple pedestrians, resulting in a total of over 300 target pedestrians. The distribution provides insights into the number of annotations for the "Cross" and “Not Cross” groups, as well as the level of disagreement among annotators concerning each target pedestrian.

Moreover, we observed that annotators tended to identify different key moments as triggers for changing their estimations of the crossing intention of target pedestrians. This implies that annotators believed the crossing intention of pedestrians in the present moment could change at various timings. Additionally, the crossing intention of certain pedestrians exhibited shifts during their interaction with vehicle drivers. To illustrate this variability, Figure 3(b) depicts the distribution of key moments (sequential frame numbers) for each target pedestrian, where intention annotations changed according to different annotators. The results demonstrate that annotators identified these key moments differently, and some pedestrians’ crossing intentions indeed changed over the course of their interaction with vehicle drivers.

3 Experimental Results

3.1 Reasoning with Large-scale Multimodal Models (LMM)

To highlight the importance of the reasoning annotations provided in the PSI dataset, we conducted a series of experiments on pedestrian intent prediction and driving decision estimation using several state-of-the-art large-scale multimodal AI models (LMM). Selected results are presented in Figure 4. The AI models were prompted with the same instructions given to human annotators, guiding both estimation and reasoning processes. Both the models and human annotators were encouraged to justify their estimations from multiple perspectives, including the behavior of other road users, road layout and traffic signals, and human interaction norms. Since the evaluated models do not support video input, we provided a sequence of keyframes from each video and asked the models to perform estimation and reasoning based for the final frame.

Pedestrian Intent and Driving Decision. From the results, we observe that the performance of all evaluated models is inconsistent, particularly in complex urban scenarios or when the target pedestrian is positioned far from the ego-vehicle, as shown in the rightmost two columns of Figure 4. Furthermore, we find that the models’ estimations of pedestrian crossing intent and the corresponding driving decision are often not aligned. For example, in the first and third columns, Gemini 2.5 Pro [22] expresses uncertainty about whether the pedestrian will cross in front of the ego-vehicle, yet

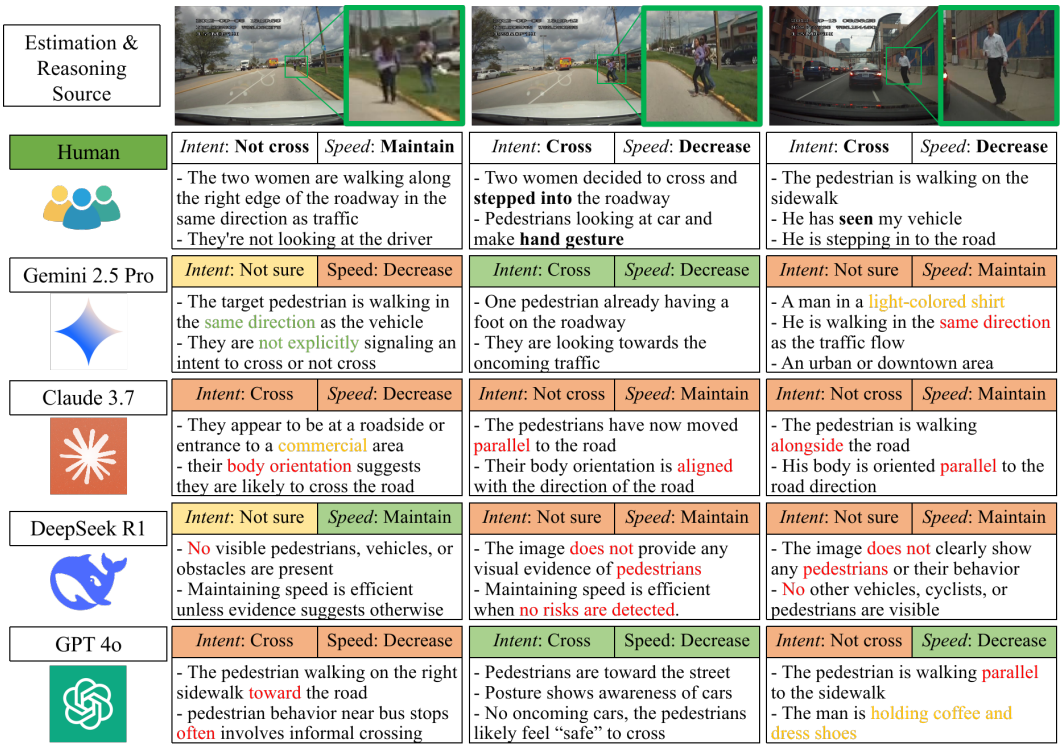


Figure 4: Comparison of pedestrian intent, driving decision estimations, and reasoning between human annotators and state-of-the-art AI models, all given the same prompts. For intent and driving decision: green indicates correct predictions matching human annotations, orange denotes incorrect predictions, and yellow marks differing but reasonable predictions. For reasoning: green highlights relevant and accurate justifications, red indicates incorrect or flawed reasoning, and yellow shows irrelevant information.

Table 3: Comparison of Reasoning of Pedestrian Intent Prediction using LMM

Model	CIDEr ↑	BLEU-4 ↑	ROUGE ↑	METEOR ↑	BERTScore ↑
Qwen-VL (7B) [3]	0.0216	0.0272	0.1786	0.1891	0.8557
InternVL (8B) [12]	0.0230	0.0337	0.1829	0.1806	0.8548
SmoVLM (2.2B) [39]	0.0227	0.0275	0.1797	0.1900	0.8558
LLaMA 3.2 Vision (11B) [18]	0.0194	0.0250	0.1935	0.1398	0.8625

it produces two different driving decisions in response-highlighting a lack of internal consistency in reasoning.

Reasoning Generation. Further challenges arise when the evaluated models attempt to generate reasoning to justify their estimations. One major issue is inaccurate perception of pedestrians and other small targets. All tested models struggled to determine the direction the target pedestrian was moving, which is a critical cue for intent prediction. DeepSeek R1 [23], in particular, failed most tasks, often unable to detect any pedestrians in the scene, resulting in randomly generated estimations and explanations based purely on general knowledge rather than visual evidence. Additionally, the models frequently produced overly verbose contextual descriptions, often fixating on irrelevant visual details. For example, GPT-4o [40] remarked that a pedestrian was “holding coffee,” while Gemini 2.5 Pro noted the individual was “in a light-colored shirt.” Although factually correct, such observations are not informative for the task of pedestrian intent or driving decision estimation. In some cases, models also exhibited hallucinations, relying on pre-trained priors instead of the actual input images. For instance, in the first column, GPT-4o correctly noted that the pedestrian was walking “toward the road” but still predicted they will “cross”, justifying this with the assumption that “pedestrian behavior near bus stops often involves informal crossing”, which is an inference not clearly supported by the visual context.

Moreover, we conducted a reasoning generation task for pedestrian intent estimation using the PSI dataset. We evaluated several pre-trained vision-language models (VLMs), including InternVL,

Table 4: Comparison of intention prediction results (%)

Baseline	Input	Acc	F1	$\text{Acc}_{\mathbf{avg}}$
LSTM [26]	box	$51.49_{\pm 0.84}$	$48.88_{\pm 1.42}$	$54.02_{\pm 0.91}$
Transformer [49]	box	$56.66_{\pm 0.44}$	$50.77_{\pm 1.18}$	$53.91_{\pm 0.39}$
PIE [43]	box+video	$59.04_{\pm 0.14}$	$56.13_{\pm 0.09}$	$61.93_{\pm 0.08}$
LSTM+*disagr.*	box	$52.75_{\pm 0.12}$	$49.60_{\pm 0.09}$	$56.74_{\pm 0.15}$
Transformer+*disagr.*	box	$59.04_{\pm 1.11}$	$54.67_{\pm 0.38}$	$59.67_{\pm 0.45}$
PIE+*disagr.*	box+video	$59.49_{\pm 0.07}$	$56.50_{\pm 0.18}$	$62.26_{\pm 0.15}$
eP2P	box+video	$60.30_{\pm 0.18}$	$57.01_{\pm 0.08}$	$62.26_{\pm 0.11}$
eP2P+*disagr.*	box+video	$61.67_{\pm 0.07}$	$57.92_{\pm 0.05}$	$62.39_{\pm 0.03}$
eP2P+*disagr.*+*rsn*	box+video	$\mathbf{63.09}_{\pm 0.16}$	$\mathbf{58.55}_{\pm 0.03}$	$\mathbf{62.53}_{\pm 0.03}$

Qwen2.5-VL, and other baselines, on the test set. Each model received the same image and a standardized prompt generated by ChatGPT-4o, aiming to predict the pedestrian's crossing intent and generate corresponding reasoning. We assessed performance using standard language generation metrics—CIDEr, BLEU-4, and ROUGE—providing a quantitative comparison across models and report the results in Table 3. These results establish baseline capabilities of current VLMs and highlight the challenges of generating accurate, context-aware reasoning for pedestrian intent.

These findings underscore the critical and complementary role of textual reasoning annotations provided by human annotators. Human-generated explanations are consistently more accurate, concise, and focused, avoiding distractions from irrelevant details and instead emphasizing subtle yet crucial visual cues and social interaction norms. Importantly, such reasoning often reflects human social understanding that goes beyond direct scene description. For example, in the second column, annotators inferred the pedestrian's intent to cross based on subtle behaviors, such as making eye contact with the driver and gesturing, demonstrating the depth of natural reasoning. This highlights the central motivation of our work: leveraging human-like reasoning to complement visual input and improve the reliability and interpretability of pedestrian intent and driving decision estimation.

3.2 Pedestrian Intent and Trajectory Prediction

Given the proposed dataset, we will perform two tasks including pedestrian intent prediction and pedestrian trajectory prediction, to demonstrate the potential tasks that can be completed with the dataset and the use cases.

Explainable Pedestrian Trajectory Prediction (eP2P). We developed an explainable Pedestrian Trajectory Prediction model to complete multiple tasks, and the model consists of two modules, *Pedestrian Situated Intent Prediction* and *Trajectory Distribution Prediction*. The intent prediction module consists of a visual encoder and a text encoder to extract the corresponding features from the input video and textual reasoning annotations, as well as a global-local context feature fusion network to aggregate the global information of the whole scene and local cues from the region near the target pedestrian. Besides, an intent predictor and a reasoning generator are deployed to predict the crossing intent and the corresponding reason for the prediction, respectively. For the trajectory prediction module, an encoder and decoder are used to integrate the spatial-temporal information of the observed moving location of the pedestrian and the crossing intent estimation. Subsequently, the trajectory predictor predicts the future locations of the target pedestrian, while the uncertainty predictor estimates hyper-parameters for the model uncertainty.

Experiments Setup and Metrics. All 196 videos in the PSI dataset are split into three subsets, training/validation/test, by $0.75 : 0.05 : 0.2$. We sample clips of frame length 60 with an overlap ratio of 0.9[43]. For intention prediction, with 0.5 second (15 frames) observation as input, we predict the target pedestrian intention at the 16^{th}-frame. For trajectory prediction, with the observed 0.5 second sequence, we predict the target pedestrian's trajectory distributions for the following 0.5, 1.0, and 1.5 seconds. For intent prediction, we report the overall accuracy (Acc), *macro* average F1 score (F1), and class-wise average accuracy ($\text{Acc}_{\mathbf{avg}}$). In particular, the class-wise average accuracy is computed as the average accuracy across both "corss" and "not cross" cases. Contrasted with the overall accuracy, denoted as Acc, $\text{Acc}_{\mathbf{avg}}$ offers a more robust evaluation of model performance, particularly in scenarios where the test data exhibits imbalance. For trajectory prediction, we evaluate the trajectory prediction using the metrics ADE/FDE and C_{ADE}/C_{FDE}.

More details of the proposed framework, experiments setup and implementation can be found in the *supplementary* material.

Table 5: Comparison of errors (pixels) of different baselines on PSI datasets for pedestrian trajectory prediction

Method	0.5s				1.0s				1.5s			
	$ADE\downarrow$	$FDE\downarrow$	$C_{ADE}\downarrow$	$C_{FDE}\downarrow$	$ADE\downarrow$	$FDE\downarrow$	$C_{ADE}\downarrow$	$C_{FDE}\downarrow$	$ADE\downarrow$	$FDE\downarrow$	$C_{ADE}\downarrow$	$C_{FDE}\downarrow$
LSTM [26]	31.65	54.75	21.54	37.71	57.80	110.90	39.82	76.96	86.98	176.74	60.18	122.69
PIE [43]	24.26	33.97	15.55	22.31	35.80	62.35	23.53	41.86	52.61	109.45	35.15	74.54
BiTraP-D [59]	25.34	31.57	16.67	20.83	35.14	60.37	23.32	40.71	52.13	111.58	35.11	76.49
SGNet [52]	24.08	40.22	15.32	26.74	43.72	85.18	29.03	57.81	67.38	143.01	45.41	97.67
SGNet+*evi* [52]	22.24	38.93	14.10	25.63	41.12	79.85	27.07	53.71	63.49	136.19	42.46	92.52
BiTraP-NP [59]	19.11	27.83	12.38	18.37	31.91	60.18	21.17	40.71	48.37	102.54	32.52	69.74
eP2P	**18.91**	**27.38**	**10.87**	**16.68**	**30.02**	**55.48**	**18.50**	**36.05**	**45.55**	**97.00**	**29.24**	**64.85**

3.2.1 Results Comparison

Pedestrian Situated Intent Prediction. The results of pedestrian situated intent prediction are reported in Table 4. We compare our model with several baselines including LSTM [26], Transformer [49], and PIE [43] with different inputs and annotations to calculate the learning objectives. Specifically, the baselines LSTM and Transformer only take the bounding boxes of the observed target pedestrian as input, while PIE leverages the local visual images in addition. The baselines are trained using the binary cross-entropy loss. In contrast, our proposed model incorporates both global and local observations as input knowledge and constrains the model training with both intent and reasoning annotations. We also compare the impact of utilizing the disagreement among all annotators to mitigate the effects of ambiguous situations.

Comparing PIE with the LSTM and Transformer baselines, we can observe how the visual observation of the local region, including the appearance of the target pedestrian, can enhance intent prediction accuracy significantly. When contrasting our **eP2P** model with PIE, we notice a performance improvement, F1 score increases from 56.13% to 57.01%, due to the fusion of global-local contextual knowledge. Moreover, by incorporating the disagreement score among annotators to reweigh the learning objectives and mitigate distractions from uncertain cases, we observe further improvement in accuracy by 1.3%. Finally, integrating the reasoning generator module and training the model with textual annotations enhances the intent prediction accuracy by 1.4%, and the explanation generation module contributes to better human comprehension and verification.

Trajectory Prediction. In Table 5, we compare our model with the baseline LSTM, which only inputs bounding boxes of the observed trajectory of the target pedestrian as input. Additionally, BiTraP estimates the long-term goal (end-point) of trajectories and introduces a novel bi-directional decoder to improve longer-term trajectory prediction accuracy, while SGNet estimates and uses goals at multiple temporal scales of trajectories. From the results, we observe that our model outperforms the compared baselines on all tasks, demonstrating the contribution of intent estimation and evidential uncertainty prediction to the trajectory prediction task. Moreover, we incorporate the evidential uncertainty prediction layer and corresponding training objectives into the SGNet framework and train the model with the same data, resulting in a boost in trajectory prediction accuracy (ADE@1.5s from 57.81 to 53.71). Such results further justify the effectiveness of the evidential uncertainty prediction design.

4 Conclusion

This paper introduced the Pedestrian Situated Intent (PSI) dataset, which captures pedestrian crossing intentions from an autonomous vehicle's perspective, with a focus on temporal and spatial context. A key innovation of PSI is its inclusion of human-annotated reasoning explanations, which provide deep insights into the cognitive processes behind pedestrian intentions and driver decision-making. These annotations go beyond mere behavior prediction, enabling models that not only forecast actions but also align with human reasoning. PSI supports critical tasks like intention prediction, trajectory forecasting, and reasoning generation, advancing the development of socially aware autonomous systems that can reason and act in a manner consistent with human thought processes. The dataset and its annotations are publicly available to drive further research and innovation in the field.

5 Acknowledgement

This project was funded by the Toyota Collaborative Safety Research Center. This material is also based upon work supported by the National Science Foundation under Grant No.2145565.

References

[1] Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, 1991.

[2] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Viena2: A driving anticipation dataset. 2018.

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[4] Andrew Best, Sahil Narang, Daniel Barber, and Dinesh Manocha. AutonoVi: Autonomous vehicle planning with dynamic maneuvers and traffic constraints. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2629–2636. IEEE, 2017.

[5] Apratim Bhattacharyya, Daniel Olmeda Reino, Mario Fritz, and Bernt Schiele. Euro-pvi: Pedestrian vehicle interactions in dense urban centers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2021.

[6] Alexandra M Boggs, Behram Wali, and Asad J Khattak. Exploratory analysis of automated vehicle crashes in california: A text analytics & hierarchical bayesian heterogeneity-based approach. *Accident Analysis & Prevention*, 135:105354, 2020.

[7] Ruthvik Bokkasam, Shankar Gangisetty, AH Abdul Hafez, and CV Jawahar. Pedestrian intention and trajectory prediction in unstructured traffic using idd-ped. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 763–770. IEEE, 2025.

[8] Michael Bratman. Two faces of intention. *The Philosophical Review*, 93(3):375–405, 1984.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Hao Chen and Xi Zhang. Path planning for intelligent vehicle collision avoidance of dynamic pedestrian using att-lstm, msfm and mpc at un-signalized crosswalk. *IEEE Transactions on Industrial Electronics*, 2021.

[11] Tina Chen and Renran Tian. A survey on deep-learning methods for pedestrian behavior prediction from the egocentric view. In *IEEE International Intelligent Transportation Systems Conference*, pages 1898–1905. IEEE, 2021.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[13] Philip R Cohen and Hector J Levesque. Intention is choice with commitment. *Artificial intelligence*, 42(2-3):213–261, 1990.

[14] Seyedehsan Dadvar and Mohamed M Ahmed. California autonomous vehicle crashes: Explanatory data analysis and classification tree. Technical report, 2021.

[15] Debargha Dey and Jacques Terken. Pedestrian interaction with vehicles: Roles of explicit and implicit communication. In *Proceedings of the International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 109–113, 2017.

[16] Joshua E Domeyer, John D Lee, and Heishiro Toyoda. Vehicle automation–other road user communication and coordination: Theory and mechanisms. *IEEE Access*, 8:19860–19872, 2020.

[17] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024.

[19] Md Fazle Elahi, Taotao Jing, Zhengming Ding, and Renran Tian. Mindread: Enhancing pedestrian-vehicle interaction with micro-level reasoning data annotation. *International Journal of Human–Computer Interaction*, pages 1–16, 2024.

[20] Md Fazle Elahi, Jithesh Gugan Sreeram, Xiao Luo, and Renran Tian. A novel adaptation of information extraction algorithm to process natural text descriptions of pedestrian encounters. In *IEEE International Intelligent Transportation Systems Conference*, pages 1906–1912. IEEE, 2021.

[21] Harshayu Girase, Haiming Gang, Srikanth Malla, Jiachen Li, Akira Kanehara, Karttikeya Mangalam, and Chiho Choi. LOKI: Long term and key intentions for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9803–9812, 2021.

[22] Google. Gemini. Large language model, 2025. Accessed October 22, 2025.

[23] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[24] Ke Guo, Wenxi Liu, and Jia Pan. End-to-end trajectory distribution prediction based on occupancy grid maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2242–2251, 2022.

[25] Michael Herman, Jörg Wagner, Vishnu Prabhakaran, Nicolas Möser, Hanna Ziesche, Waleed Ahmed, Lutz Bürkle, Ernst Kloppenburg, and Claudius Gläser. Pedestrian behavior prediction for automated driving: Requirements, metrics, and relevant features. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[27] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision*, pages 584–600. Springer, 2020.

[28] Jia Huang, Peng Jiang, Alvika Gautam, and Srikanth Saripalli. Gpt-4v takes the wheel: Promises and challenges for pedestrian behavior prediction. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 134–142, 2024.

[29] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2019.

[30] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision*, pages 234–251, 2018.

[31] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European Conference on Computer Vision*, pages 563–578, 2018.

[32] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2):1940–1947, 2019.

[33] Iuliia Kotseruba, Amir Rasouli, and John K Tsotsos. Joint attention in autonomous driving (JAAD). *arXiv preprint arXiv:1609.04741*, 2016.

[34] Kunming Li, Stuart Eiffert, Mao Shan, Francisco Gomez-Donoso, Stewart Worrall, and Eduardo Nebot. Attentional-gcnn: Adaptive pedestrian trajectory prediction towards generic autonomous vehicle use cases. In *IEEE International Conference on Robotics and Automation*, pages 14241–14247. IEEE, 2021.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[36] Bingbin Liu, Ehsan Adeli, Zhangjie Cao, Kuan-Hui Lee, Abhijeet Shenoi, Adrien Gaidon, and Juan Carlos Niebles. Spatiotemporal relationship reasoning for pedestrian intent prediction. *IEEE Robotics and Automation Letters*, 5(2):3485–3492, 2020.

[37] Tong Liu, Akash Venkatachalam, Pratik Sanjay Bongale, and Christopher Homan. Learning to predict population-level label distributions. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 1111–1120, 2019.

[38] Srikanth Malla, Behzad Dariush, and Chiho Choi. Titan: Future forecast using action priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11186–11196, 2020.

[39] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, et al. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

[40] OpenAI. ChatGPT. Large language model, 2025. Accessed October 22, 2025.

[41] Hannah RM Pelikan. Why autonomous driving is so hard: The social dimension of traffic. In *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, pages 81–85, 2021.

[42] Amir Rasouli and Iuliia Kotseruba. Diving deeper into pedestrian behavior understanding: Intention estimation, action prediction, and event risk assessment. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 1381–1388. IEEE, 2024.

[43] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K Tsotsos. PIE: A large-scale dataset and models for pedestrian intention estimation and trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6262–6271, 2019.

[44] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. Understanding pedestrian behavior in complex traffic scenes. *IEEE Transactions on Intelligent Vehicles*, 3(1):61–70, 2017.

[45] Paschal Sheeran and Thomas L Webb. The intention–behavior gap. *Social and personality psychology compass*, 10(9):503–518, 2016.

[46] Matus Sucha, Daniel Dostal, and Ralf Risser. Pedestrian-driver communication and decision strategies at marked crossings. *Accident Analysis & Prevention*, 102:41–50, 2017.

[47] Dapeng Tao, Jun Cheng, Zhengtao Yu, Kun Yue, and Lizhen Wang. Domain-weighted majority voting for crowdsourcing. *IEEE transactions on neural networks and learning systems*, 30(1):163–174, 2018.

[48] Renran Tian, Lingxi Li, Kai Yang, Stanley Chien, Yaobin Chen, and Rini Sherony. Estimation of the vehicle-pedestrian encounter/conflict risk on the road based on tasi 110-car naturalistic driving data collection. In *IEEE Intelligent Vehicles Symposium Proceedings*, pages 623–629. IEEE, 2014.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[50] Erik Vinkhuyzen and Melissa Cefkin. Developing socially acceptable autonomous vehicles. In *Ethnographic Praxis in Industry Conference Proceedings*, volume 2016, pages 522–534. Wiley Online Library, 2016.

[51] Shaun Wallace, Tianyuan Cai, Brendan Le, and Luis A Leiva. Debiased label aggregation for subjective crowdsourcing tasks. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–8, 2022.

[52] Chuhua Wang, Yuchen Wang, Mingze Xu, and David J Crandall. Stepwise goal-driven networks for trajectory prediction. *IEEE Robotics and Automation Letters*, 7(2):2716–2723, 2022.

[53] Jing Wang and Xin Geng. Classification with label distribution learning. In *IJCAI*, volume 1, page 2, 2019.

[54] Wenjing Wu, Hongfei Jia, Qingyu Luo, and Zhanzhong Wang. Dynamic path planning for autonomous driving on branch streets with crossing pedestrian avoidance guidance. *IEEE Access*, 7:144720–144731, 2019.

[55] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022.

[56] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2174–2182, 2017.

[57] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9523–9532, 2020.

[58] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith A Redmill, and Ümit Özgüner. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles*, 7(2):221–230, 2022.

[59] Yu Yao, Ella Atkins, Matthew Johnson-Roberson, Ram Vasudevan, and Xiaoxiao Du. Bitrap: Bi-directional pedestrian trajectory prediction with multi-modal goal estimation. *IEEE Robotics and Automation Letters*, 6(2):1463–1470, 2021.

[60] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018.

[61] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. Explainability of vision-based autonomous driving systems: Review and challenges. *arXiv preprint arXiv:2101.05307*, 2021.

[62] Shile Zhang, Mohamed Abdel-Aty, Yina Wu, and Ou Zheng. Pedestrian crossing intention prediction at red-light using pose estimation. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions to the PSI dataset and experiments in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed the challenges of the data collection stage and the limitations of the dataset, as well as the limitations caused from existing AI models in our experiments in the manuscript.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This is a dataset/benchmark work which does not have theory assumptions and proofs. Important references and motivations are properly cited.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the dataset and processing code, experimental code and instructions are provided in the manuscript or open-source.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the dataset is host on HuggingFace, and corresponding codes are open-source on Github.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In our paper, we detailedly described how the experimental results are obtained, and more codes including training and test process are provided on Github.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: We demonstrate the contribution of the new PSI dataset in this dataset/benchmark track paper, thus no statistical significance analysis of the experimental results are conducted.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments involved in this paper are based on public AI models such as GPT 4o or DeepSeek R1. Thus people can reproduce the results using those tools.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that we follow the NeurIPS Code of Ethics in the whole process of this research paper.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both positive and potential negative societal impact caused by the limitations of the PSI dataset is discussed.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: All data collected in this dataset are thoroughly reviewed by legal teams and obtained necessary consent.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All referenced or reused tools and codes are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Detailed documents and codes are provided alongside the dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Detailed instructions and data collection details are provided in the paper and open source assets.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This work is conducted with Institutional Review Board approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The usage of LLM for writing polish and experiments are clearly mentioned.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.